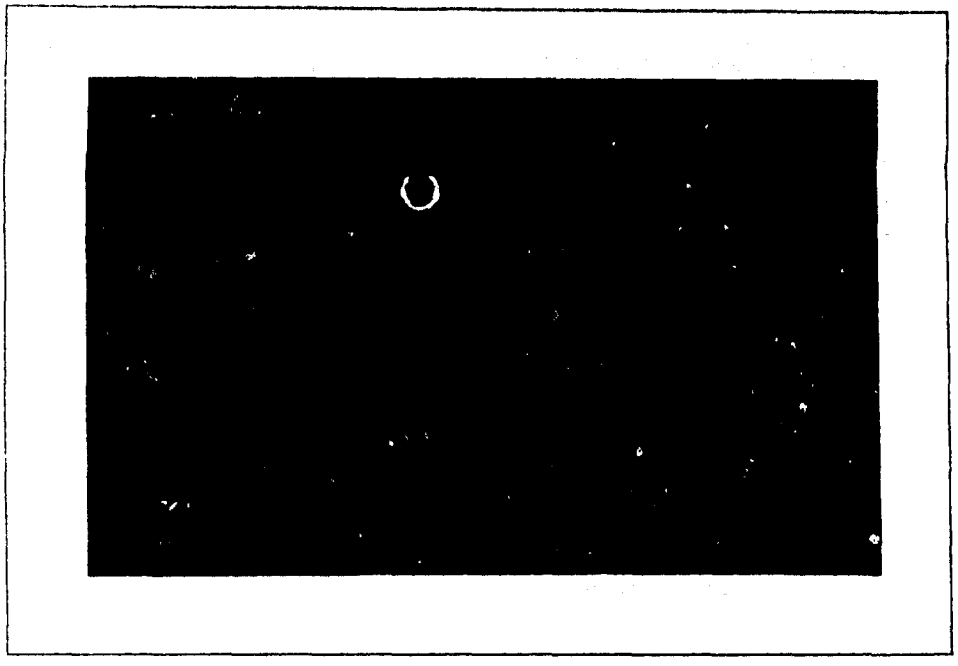


The Artificial Intelligence and Psychology Project

Departments of
Computer Science and Psychology
Carnegie Mellon University

Learning Research and Development Center
University of Pittsburgh

2

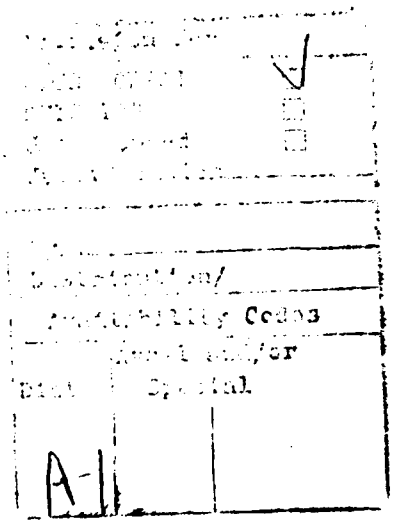
COGNITIVE ARCHITECTURES AND RATIONAL ANALYSIS: COMMENT

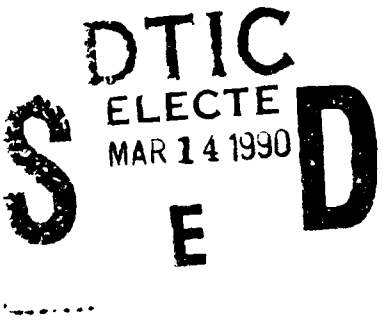
Technical Report AIP - 58

Herbert A. Simon

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

17 March 1989





This research was supported in part by the Defense Advanced Research Projects Agency, Department of Defense, ARPA Order 3597, monitored by the Air Force Avionics Laboratory under contract F33615-81-K-1539 and the Computer Sciences Division, Office of Naval Research, under contract number N00014-86-K-0678. Reproduction in whole or part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b RESTRICTIVE MARKINGS
2a SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT: Approved for public release; Distribution unlimited
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): AIP - 58	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION: Carnegie-Mellon University; 6b OFFICE SYMBOL (If applicable)	7a NAME OF MONITORING ORGANIZATION: Computer Sciences Division, Office of Naval Research
6c ADDRESS (City, State, and ZIP Code): Department of Psychology, Pittsburgh, Pennsylvania 15213	7b ADDRESS (City, State, and ZIP Code): 800 N. Quincy Street, Arlington, Virginia 22217-5000
8a NAME OF FUNDING / SPONSORING ORGANIZATION: Same as Monitoring Organizatior; 8b OFFICE SYMBOL (If applicable)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-86-K-0678
8c ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS p4000ub201/7-4-86

PROGRAM ELEMENT NO	PROJECT NO	TASK NO.	WORK UNIT ACCESSION NO
N/A	N/A	N/A	N/A

11 TITLE (Include Security Classification)

Cognitive Architectures and Rational Analysis: Comment

12 PERSONAL AUTHOR(S) Herbert A. Simon

13a TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical	FROM 86Sept15 TO 91Sept14	1989 March 17	21

16 SUPPLEMENTARY NOTATION

To appear in VanLehn, K. (Ed.) ARCHITECTURES FOR INTELLIGENCE. Erlbaum: Hillsdale, NJ

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)

cognitive architectures: rationality, optimization, cognitive processes, adaptation

19 ABSTRACT (Continue on reverse if necessary and identify by block number)

SEE REVERSE SIDE

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	

22a NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Dr. Alan L. Meyrowitz	(202) 696-4302	N00014

DD FORM 1473, 84 MAR 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

ABSTRACT

John Anderson has written a provocative chapter whose thesis may be oversimplified to read: To understand the behavior of an adaptive organism, don't study the organism; study its environment.

To claim that architecture is more notation than substance is to make the same claim for the brain -- the fact that it supports adaptive behavior makes unnecessary any curiosity about how it operates. The exact way in which neurons accomplish their functions is not important -- only their functional capabilities and the organization of these. Nothing else will show through to "behavior."

But what does show through is precisely what we have been calling "architecture." And for that reason architecture is by no means all notation; it has real substance in its effects on behavior. In my view, Anderson assigns too little weight to architecture (and by implication to strategies) as determinants of adaptive behavior.

to FLD 18

Cognitive Architectures and Rational Analysis: Comment

Herbert A. Simon
Carnegie-Mellon University

John Anderson has written a provocative chapter whose thesis may be oversimplified to read:

> **To understand the behavior of an adaptive organism, don't study the organism; study its environment.**

When we are served a gelatin desert shaped like an exotic animal, and want to know why it takes that form, we are well advised to look for the mold in which it was made. We need to know nothing about gelatin except that, in liquid form it adapts its shape to the shape of whatever vessel receives it, and retains that shape when it sets. The adaptive principle of gelatin (its "goal") when in liquid form is to minimize the level of its center of gravity.

Anderson invites us to view the human mind, when confronted with a new kind of task, as a bowl of gelatin that fits itself optimally to the demands of the task, acquiring the strategy that is most appropriate for performing it efficiently. If we accept his invitation, then we are relieved of the tedium (if that is the right word) of comparing cognitive architectures and experimenting to determine which one best describes human task behavior. We are even excused from the tedium of inventing the architectures in the first place: what goes on in the mind doesn't matter (much).

The Optimal Adaptation Hypothesis

The hypothesis that adaptive systems are to be understood by computing their optima from the structure of their environments has played a central role in several domains of science: most notably in economics and evolutionary theory. As we shall see, it has not been absent from psychology. It may be instructive, and useful for

evaluating Anderson's proposal, to see how it has fared in these various domains. I begin with economics.

Optimization in Economics

The foundation stone of contemporary neo-classical economics is the hypothesis that economic actors are rational (i.e., perfectly adaptive), and that their rationality takes the form of maximizing expected subjective utility. The term "maximizes" is clear enough. "Expected" is meant in its technical sense in probability theory -- that is, as in the phrase "expected value." "Subjective" means "in terms of the actor's estimates of the relevant probabilities of events. "Utility" means the actor's own ordering of preferences among outcomes, assumed to be consistent, but otherwise wholly arbitrary.

Now how can we use this formula to predict behavior? Let us, for example, suppose that a businessman is faced with a demand schedule for the product of his firm (which tells him the amount he can sell at any given price) and a cost schedule (which tells how much it will cost him to produce any given amount). As yet we can conclude nothing about his behavior unless we know what he is trying to accomplish -- what his utility function is. *If* we make the additional assumption that his utility is measured by his net profit, then we can instantly predict the quantity he will produce (the quantity that maximizes the difference between total revenue and total cost).[1]

Difficulties with Economic Optimization

That sounds rather powerful and convenient. A description of the environment (the demand and cost schedules), and an innocent assumption about motives (utility = profit) is all we need to know to predict behavior. No tiresome inquiries into the

---

[1] I have made this businessman masculine because he is just a caricature of a real human being, neither his sex nor his thought processes affecting his behavior.

businessman's mental states or processes.

Defining Goals. Complications only begin to arise when we try to apply this model to the real world. How do we know the real businessperson wants to maximize profit? Perhaps he or she wants to maximize the respect received from the community, or the friendliness of workers or of customers. That would lead to very different behavior.

If we think that profit really is the criterion of choice, it cannot be because of any requirement imposed by the task, but because of something stored in the businessperson's head. The content and shape of the utility function can only be determined by empirical study of what goes on in that head. No amount of study of the task environment will help.

We may wish to dismiss this objection as not very fundamental. After all, if we are talking about cognitive tasks, isn't it reasonable to assume, without a great deal of fuss about empirical verification, that people want to perform such tasks as quickly, accurately, and efficiently as possible? For many purposes of experimentation and explanation that's a good enough assumption (though we may have to worry a bit about the speed/accuracy tradeoff), and we make it all the time in our laboratory work.

Incomplete Information. But there are other difficulties, even if we ignore this one. Does the businessperson really know the cost and demand schedules, and know them accurately? When we observe the actual operation of businesses, we find that a great deal of time and effort is devoted to finding out just what these schedules are -- how much it will cost to produce the product and how much of it can be sold at different prices. And seldom do businesspeople imagine that they really know the answers to these questions with any accuracy. The real world is filled with large quantities of uncertainty.

At this point the estimates of expected utilities come into play. Where do the subjective probabilities come from? Can we infer these from the task environments, or must we enter the heads of the actors to see how they estimate them (or whether, in fact, they do or don't use anything resembling subjective probabilities in their deliberations)? So in a world where the givens are not really given but must be inferred by the adaptive organism, there is in fact no way in which behavior can be predicted from the external environment without consideration of how that environment is apprehended (and with what accuracies and errors) by the intendedly adaptive actors.

Generation of Representations and Alternatives. Other limitations of the theory of maximization of subjective expected utility are equally important. The theory also assumes the formulation of the problem to be given to the actors, as well as a complete inventory of possible alternatives of action. There is no place in such a theory for a focus of attention or for a search for new alternatives (e.g., for new products, for new markets). The problem representation and the alternatives, like the cost and demand schedules, are assumed to be given as an objective part of the external environment. These heroic assumptions create a further gulf between the optimization theory and the conditions under which real world economic decisions are made.

I have made all of these points elsewhere at much greater length (Simon, 1982, Vol. 2, Secs. VII and VIII), and many others have made them also. Economists today are conscious of a crisis in their discipline that derives directly from these difficulties; but this is not the place to make the case against neo-classical economics, nor to prescribe for its ills.

Auxiliary Assumptions. Before leaving the economic example, however, one other point should be made. In order to restore at least a modicum of realism to

their models, and to fit them to the observed phenomena, economists are accustomed to introducing into their reasoning various qualifying assumptions that amount to departures from the model of perfect adaptation. For example, the Keynesian explanation of unemployment hinges essentially on the assumption that labor is "sort of" rational, but not wholly rational. Labor wants higher wages -- but fails to distinguish accurately between increases in money wages and increases in real wages. Since it is real wages that buy bread and meat, and not money wages, this failure in discrimination is a genuine irrationality, or imperfect adaptation. Yet it is precisely such grains of irrationality in the economic oyster that produce the pearls of the real-world phenomena -- in this case unemployment.

The plain fact is that the conclusions that economists draw from their optimizing models seldom depend critically upon the optimizing assumptions, but they *do depend critically upon the auxiliary assumptions of departures* from rationality -- that is, assumptions of imperfect adaptation. It is these qualifying auxiliary assumptions that do the greatest part of the work in leading economic theories to their conclusions. Change the assumptions, while retaining the postulate of optimization, and you change the conclusions.

Once a behavior has been observed, it is usually rather easy to find some auxiliary assumptions describing circumstances under which the observed behavior is the optimal response. It is especially easy to find such assumptions if we do not require that they be supported by any direct empirical evidence of their presence. And it is even easier if we may introduce new auxiliary assumptions for each new phenomenon we wish to explain. Precautions need to be taken against this kind of adhokery not only in the behavioral sciences, but in the natural sciences as well. As Poincare put it, when discussing some experimental anomalies that arose in physics at the turn of the century: "An explanation was necessary, and was forthcoming;

they always are. hypotheses are what we lack the least."

In economics, then, the apparent escape from the study of the psychology of economic actors is illusory. It is not, in fact, possible to predict their behavior solely, or even mainly, from a study of their task environments. Economists have avoided this unpleasant conclusion because they have been willing to make their auxiliary assumptions -- their assumptions about human thought processes -- from the armchair without requirements of empirical verification. That this kind of "casual empiricism" (the economists' name for it) is rampant in economics is easily verified by examining the professional literature. It is hardly a model for the study of human rationality by other disciplines.

Explanation by means of optimization assumptions is sometimes extolled as a method of great parsimony. But when it is combined with the lavish use of auxiliary assumptions it is far from parsimonious. Unless the phenomena that are explained are numerous and complex, the degrees of freedom provided by the ad hoc assumptions can easily outnumber the data points to be explained.

Adaptation by Natural Selection

Economics is not the only discipline in which optimization theory plays a central role. In the Darwinian theory of evolution by natural selection, appeal is often made to arguments of optimization. It is sometimes asserted that, by the terms of the theory, only systems that optimize will survive. This assertion has also often been borrowed by neoclassical economists to explain why they need not be concerned with the processes of decision: whatever the processes, only the firms that maximize profits will survive.

When we examine the logic of natural selection more closely, we see that it does not imply optimization at all. All it implies is that if there are two or more organisims competing for the same resources (occupying the same ecological niche),

the one that uses the resources most efficiently. In terms of multiplying its numbers (fitness) will replace the others. It does not have to be best in any absolute sense it need only be better than its competitors.

To use evolutionary theory to predict behavior, we must first know what behaviors are available, for the theory makes no claim that all the relevant behaviors have been discovered or are known to the actors. Before it was opened to European settlement, the biota of Australia had perhaps reached some sort of evolutionary equilibrium. That this equilibrium was not in any sense a global optimum was revealed as soon as rabbits were introduced and showed by multiplying that they were "fitter" than some of the indigenous organisms.

The theory of natural selection is not an optimizing theory for two reasons. First, it can, at best, produce only local optima, because it works by hill-climbing up the nearest slope. It has no mechanism for jumping from peak to peak, hence is likely to be trapped repeatedly on knolls well below the highest summit. Second, it selects only among the alternatives that are available to it, and has no way of guaranteeing that new, and better, alternatives, will not appear from time to time. Indeed, the whole story of biological evolution is a tale of just such successive appearances. So evolution deals with systems that never reach an optimum, and whose definition of the optimum is continually changing.

It is not surprising, therefore, that we see very little prediction in the literature of evolution. What we see are explanations, post hoc, of the observed facts. If the facts were different, we would have no great difficulty in finding another evolutionary explanation for them. There are many adaptive paths to the fitness goal.

For a further elaboration of these arguments, see my *The Sciences of the Artificial*, 2nd edition: "The Evolutionary Model," pages 52-60, which also discusses the myopia of evolution and why it doesn't produce optimization. There is a related

discussion in the chapter on "Rationality and Teleology" in my book *Reason in Human Affairs*, Chapter 2, especially on pages 66-72, where myopia is discussed again.

For our present purposes, the central point -- both with respect to evolutionary theory and economic theory that uses auxiliary assumptions -- is that "optimization" theories of these kinds do not determine unique solutions or make unique predictions. They lend themselves, therefore, to after-the-fact explanations of what we already know, but provide no convincing evidence that these explanations bear any relation to the real causal mechanisms of processes. To demonstrate causality and to understand process, we cannot limit ourselves to analyzing the environment of behavior, but must study the behaving system, and specifically the limits on its powers of adaptation.

Optimization in Psychology

The idea that adaptive behavior is to be explained by the shape of the environment also has a long history in psychology. Egon Brunswik was an early exponent of this idea (See his 1956 on the role of the distal stimulus), and it is central to Gibson's (1979) view of perception as it is of Marr's (1982). Anderson has already quoted Marr's view on this point:

> An algorithm is likely to be understood more readily by understanding the nature of the problem being solved than by examining the mechanism (and the hardware) in which it is embodied.

The evolutionary version of the optimization hypothesis, essentially identical with that set forth in standard Darwinian theory, has been espoused by Donald Campbell (1974), and another version of it is to be found in the work of Lumsden and Wilson (1981). In application to psychology, however, the argument for optimization does not have to rely on natural selection among behaviors to produce adaptation. The

organism's capabilities for learning and for solving problems provide direct means for adaptation. Of course Campbell points out that learning and problem solving processes, insofar as they employ trial and error, are themselves evolutionary mechanisms that depend on natural selection for their efficacy. An important difference, however, is that in this case the trial and error can take place in mental problem spaces, and does not necessarily rely on experimentation in the real world.

The same objections that have been raised against evolutionary explanations in biology and economics can be raised against their employment in psychology. Evolution (and learning and problem solving) do not guarantee uniqueness of the adapted behavior, much less its optimality. At best they only predict improvement over previous behavior. In no way do they provide adequate explanations of the behavior that actually occurs. Any one of many alternative local optima could be equally well explained, if they occurred, by reference to the environmental variables.

What Does Rationality Imply About Behavior?

In my paper, "The Functional Equivalence of Problem Solving Skills," reprinted in Simon (1979), Chapter 4.5, I described four strategies for solving the Tower of Hanoi problem. All four strategies are "rational," in Anderson's sense, since all four produce a solution path, from the usual starting point, of minimum length. From our knowledge of the task environment, therefore, we can predict the solution path that efficient subjects will follow, independent of their strategies and their architectures.

But what *can't we predict, without additional information about the subject?* We can't predict which strategy the subject has learned, although with the help of thinking-aloud protocols, this is ascertainable. We can't predict whether the subject, if interrupted for fifteen seconds, will be able to continue toward the solution without error -- that depends on the strategy being used. We can't predict whether the

subject will be able to solve the problem efficiently if the number of disks is increased -- that also depends upon the strategy, since different amounts of short term memory (an architectural variable, not an environmental one) are needed with different strategies.

Empirically, we find that a plurality of subjects under usual experimental conditons learn the move-pattern strategy. There is nothing in the task environment that implies that this is the strategy that will be learned. Its prevalence seems to derive from the readiness with which subjects notice sequential patterns (1-2-1-3-1-2-1-4-1-2-1-3-etc.) and remember them. This is a function of mental architecture, not of the demands of the task environment.

So while we may agree with Anderson that in this task a "signature," the solution path, can be inferred from the demands of the environment, this signature by no means exhausts our interest in the subjects' behavior. If we are interested in learning and in transfer, we need to know what strategies they adopt. Strategies are aspects of behavior that can be observed empirically without too much trouble. And the learning of strategies, as well as the consequences of employing one strategy rather than another, are functions of cognitive architecture. Since they vary with the architecture, they can be used to infer properties of the architecture.

Some Counterexamples?

Up to this point, I have been proceeding mainly on the high road of theory, trying to show why it is impossible in principle to explain adaptive behavior simply by examining the structure of the task environment. But John Anderson has provided us with three actual, and ingenious, examples of such explanation. The empiricism I have been espousing would argue that hypothetical reasoning must yield to valid counterexamples.

What are the facts? Do Anderson's models explain the phenomena he

describes as optimal responses to the task requirements? Or is the real work in his models being done by auxiliary assumptions? And even if it turns out that certain "signature" data can be predicted without much reference to architecture, do we wish to limit ourselves to these data, and are there other ways of inferring what the underlying architecture and strategies really are? Let us examine the models one by one.

The Power Law in Learning

In the first example, an optimization model is built to predict the relation between practice and speed of performance in the Seibel task, the relation being linear on a logarithmic scale. Anderson postulates a memory of independent items that are accessed through a key word index. (We see that a number of architectural assumptions, although rather general ones, have to be made at the outset. The model does not by any means derive its power *solely from knowledge* of the requirements of the task.) It is further assumed that there is a cost associated with retrieving each successive item, and a cost associated with failing to retrieve the desired item. Finally there is associated with each item a probability that the item is a desired one, and it is assumed that this probability is known, at least ordinally, by the processor. (In fact, the derivation of the log-linear relation assumes that the cardinal values of the probabilites are known.)

Probability of Retrieval. The optimization principle is to retrieve items in order of decreasing probability until the expected gain from retrieving the next item is less than the cost of retrieving it. From this principle, and using a number of auxiliary assumptions that I will mention in a moment, Anderson shows that this probability will be a linear function of the number of times the item has been used in the past.

Practice and Latency. But we still do not have a prediction of the observed relation between practice and latency. This is then obtained by the additional

derivational step that "under the transformation from need probability to latency proposed in Anderson (in press) the power function relationship remains." In the source cited, Anderson derives the transformation from the assumption that the need probabilities of items are distributed according to Zipf's law, a law that has been found empirically to fit a wide range of not unrelated phenomena. The Zipf's law assumption is quite plausible, but it is an empirical assumption, not an optimization assumption. Zipf's law is clearly doing as much of the work as is the optimization assumption.

The structure of the argument, then, is this. By use of the optimization principle with numerous auxiliary assumptions, we infer a linear relation between practice and probability of retrieval. Then, with another assumption, we derive a monotonic relation of appropriate shape between probability and latency. From this it follows that latency and practice are connected by a power law. Would it not be simpler, and more parsimonious, simply to postulate the latter law without deriving it? But whether that law holds or not is a matter of how the memory functions, and not a simple consequence of optimization.

The Auxiliary Assumptions. I must return briefly to the other auxiliary assumptions that are made in the course of the optimization argument. There appear to be about four of these. The first is embedded in Anderson's equation (2), and determines how the need probabilities change in the course of repeated memory searches -- a very specific Bayesian assumption, describing one of a large number of possible inductive rules. The second is that items are distributed by desirability accoding to the gamma function -- a very specific assumption, buttressed by some evidence from library systems. The third is that usage of items decays exponentially The fourth is that repeated retrievals of items are distributed according to the Poisson process. These last three are assumptions about the structure of the task

environment, and are empirically testable.

The conclusion that need probability varies linearly with number of uses is primarily due to the first and fourth assumptions, and is probably not very sensitive to the others. The two critical assumptions, rather than the optimization criterion, are doing most of the work of the derivation, and the first assumption is a specification of architecture or strategy, not of environment.

Notice that in the Seibel task, all needed items are actually retrieved. Hence, there is no question of stopping the search when the cost of retrieving an item is too large. It seems odd to base the derivation of the relation between frequency of past use and latency on the optimization of this inoperative mechanism.

Further Predictions of the Model. But Anderson makes additional claims of parsimony for his derivation. From it he obtains two other empirically observed effects: a forgetting function, and familiar effects of massed and distributed practice. The forgetting effect derives directly from inserting an exponential decay function among the auxiliary assumptions, and is surely derivable from the latter without any optimization. But the decay function is motivated by our knowledge that, empirically, there is memory decay, and is an assumption about architecture, not about task requirements. No parsimony here.

Likewise, the derivation of effects of spacing of practice requires two additional auxiliary assumptions -- that memory for some items decays faster than memory for others, and that the relevance of items is revived periodically. The first of these assumptions is architectural. Anderson does not show in detail how the spacing effect is derived from these assumptions, but it may be inferred that the opportunity for items to be revived over longer time intervals is at the root of it.

The explanation of the data in the Seibel task, and the others that are discussed in connection with the model, provides something less than a convincing

demonstration of the power of the model, or its parsimony. Its ability to postdict some empirical phenomena is not at all independent of architectural assumptions, and no more independent of the numerous auxiliary assumptions that have to be introduced along the way. It would appear that equally convincing post hoc characterizations of the data can be derived from these auxiliary assumptions without the aid of the optimization procedure. But these characterizations are not explanations in any deep sense. They are simply summaries of the empirically observed phenomena.

The Fan Effect

In his second example of the prediction of "signatures" from optimization in a specified task environment, Anderson uses the same model as before of the effects of practice on the need probability of items in memory. He now employs the model to show that "the fan effect is a consequence of memory using the correlation between cues and a memory's relevance to predict when the memory is needed." That is to say, people "optimize" by keeping better access to memories that are likely to be relevant.

What is notable about this example is that the prediction is quite bland, and certainly does not require anything as strong as an optimization assumption to derive it. Anderson is correct in asserting that little can be inferred from it about architecture -- *Any* architecture that kept items close at hand on the basis of recent or frequent use would do the trick. Any such architecture would exhibit the *Einstellung* effect as well!

The conclusion to be drawn is not that we can substitute optimization procedures for an understanding of architecture. Rather, it is that we need to look at more subtle phenomena -- for example, detailed thinking-aloud protocols of individual subjects -- if we want to learn about architecture and to discriminate

among different architectures. We need to detect architecture, not with a few "signatures," but with a wide range of converging phenomena. Feigenbaum and I (1984) have used just such a strategy to examine the validity of the EPAM model, which we have proposed as a (partial) architecture for human memory.

Categorization

The structure of the argument for Anderson's third example is very similar to that for the fan effect. Human beings, given repeated learning trials, gradually acquire skill in categorizing objects. If the learning time is long enough, and the set of objects not too complex, they may even learn to categorize them "optimally" by some external criterion. (The fact that many biologists since Linnaeus have spent their professional careers modifying and improving plant and animal taxonomies is perhaps evidence that in the real world optimality of systems of categories doesn't come quickly or easily.)

Anderson now shows that he can define an algorithm (a search procedure that optimizes, if at all, only asymptotically) that will improve its categorizations much as people do. Undoubtedly there are many such algorithms. What we really wish to know is which of these are actually used by people in learning to categorize. We are not interested in what categories they will ultimately learn: That is not psychology but botany or entomology or whatever science the phenomena belong to. Our interest is in the learning process itself -- not a hypothetical one or an optimal one, but the one that people use. We want a learning theory precisely because *people do not arrive at optimal classifications* immediately or costlessly. We wish to understand what sly tricks they use to arrive at clasifications at all, and what prospects there are for improvement of the process.

The example illustrates once again that the interesting and important issues for psychology are not to demonstrate that people are motivated to behave rationally

when the circumstances are simple enough to make such behavior possible. The interesting issues are precisely to determine what internal limits (computational or other) prevent people from attaining optimality, or attaining it rapidly and costlessly; and to understand how people use the computational capabilities they have in order to cope with these limits.

Optimization or Bounded Rationality?

Bounded rationality is what cognitive psychology is all about. And the study of bounded rationality is not the study of optimization in relation to task environments. It is the study of how people acquire strategies for coping with those environments; how those strategies emerge out of problem space definitions; and how built-in physiological limits shape and constrain the acquisition of problem spaces and strategies. At each of these steps there is room for alternative processes, any of which would meet satisfactorily (not optimally) the requirements of the task environment. The environment cannot predict which of these alternatives will govern the adaptive behavior.

I would repeat for cognitive psychology what I said, more than forty years ago about organization theory (*Administrative Behavior*, p. 240):

> . . . if there were no limits to human rationality administrative theory [read: cognitive theory] would be barren. It would consist of the single precept: Always select the alternative, among those available, which will lead to the most complete achievement of your goals. The need for an administrative theory [cognitive theory] resides in the fact that there *are* practical limits to human rationality, and that these limits are not static, but depend upon the . . . environment in which the individual's decision takes place."

The Two-Bladed Scissors

The moral to be drawn from our discussion is not that the task environment is unimportant in explaining the behavior of an adaptive system, but that one must

consider both the task environment and the limits upon the adaptive powers of the system. Only in the simplest cases will the system behavior be predictable from an optimization argument. Almost always, structure and limits to adaptation will, to some degree, "show through," and hence will have to be taken into account. The outer environment and the inner structure are the two blades of the scissors, and both blades must be present and operative for a satisfactory dissection of what is going on.

In writing the final, theoretical, chapter of *Human Problem Solving*, Allen Newell and I were confronted with the same problem of defining the *laws* of the behavior of adaptive, thinking, organisms. How could there be such laws if problem solvers were adaptive? Our proposed resolution of the problem can be found on pages 788 to 789 of our book. I will give its flavor by quoting a couple of passages, adding comments in square brackets.

> . . . The task environment (plus the intelligence of the problem solver) determines to a large extent the behavior of the problem solver, independently of the detailed internal structure of his information processing system. [So far we are in agreement with Anderson.]
>
> . . . A few, and only a few, gross characteristics of the human IPS are invariant over task and problem solver.
>
> These characteristics are sufficient to determine that a task environment is represented (in the IPS) as a problem space, and that problem solving takes place in a problem space. [problem space = representation.]
>
> The structure of the task environment determines the possible structures of the problem space. [The representation is not uniquely determined, but only constrained by the task environment.]
>
> The structure of the problem space determines the possible programs [strategies] that can be used for problem solving. [The strategy is not uniquely determined, but only constrained by task environment and problem space.]

The two blades of the scissors, in this formulation, are (1) the task environment and (2) the problem space conjoined with the strategy used for searching it. It is the organism that constructs a problem space and strategy to deal with the task environment. The problem space and strategy are usually adapted to the task, but

in a much weaker sense than optimization.

Among the computational limits that are important in shaping behavior are limits on the capacity of short-term memory, the presence or absence of external memory aids, the failure of human subjects to use best-first search (primarily because of short-term memory limits), and the general absence of any optimization processes for selecting problem representations.

Moreover, representations and strategies are not usually given, but have to be discovered by the problem solver. Nor are they unique; in a particular task domain, many satisfactory (not optimal) alternatives may be potentially available. Which particular representations and strategies will be discovered by particular subjects and under what circumstances cannot be predicted from a knowledge of the task structure, but depends on the other blade of the scissors as well..

The evidence that has been gathered over the past decade or more on differences in difficulty of problems that have isomorphic task domains (Kotovsky, Hayes, and Simon, 1986; Kotovsky and Simon, to be published) provides a striking example of the inadequacy of a one-bladed scissors for explaining problem-solving performance. The fact that one form of a problem can require, on average, sixteen times as much effort to solve as another isomorph of the same problem cannot be readily explained with principles of optimization. We have to discover what limitations of the problem solver prevent him or her, in the more difficult case, from proceeding in exactly the same manner as in the easier one, the two being fully isomorphic.

In those situations where behavior that is optimal can be learned, and where the behavior of the expert can properly be described as optimal, it may still be of great interest to understand how someone progresses from novice to expert performance -- that is, how learning takes place. But the very notion of "optimal" learning is ambiguous. Presumably optimal learning would be instant learning, and

since very little human learning takes place instantly, we need to understand the internal limitations to adaptation that constrain its speed.

Detecting Architectures and Strategies

I am more sanguine than is Anderson that empirical evidence can discriminate among cognitive strategies and among cognitive architectures, and can show some to be more acceptable than others as models of human information processing. But to accomplish this we have to confront the adaptive systems we are studying with complex tasks that will stretch their abilities to behave "optimally," and will reveal the structural limits on their adaptation.

It may well be that in many cases where adaptation to the external environment is reasonably effective, "signature data" will be predictable from the requirements of the situation. If people are adding up columns of figures (under circumstances where they usually get the right answers), then the sums they write down will be predictable in the manner that Anderson describes. The conclusion I would draw, however, is not that the process or the architecture of the system reaching the result is inscrutable. My conclusion would be that we need data other than these *"signature data" to discover what the process and architecture are.*

To identify mechanisms, strategic and structural, we cannot be satisfied with aggregated data of total performance (speed of performance, aggregate accuracy or success), but have to observe as many details of the ongoing processes as we can. Classical static experimental designs have to give way to studies of dynamic processes, observed through the recording of protocols, eye movements, the details of errors, and such other means as our ingenuity may suggest. It is along this route that we can solve the difficult problems of identifiability that always confront us in trying to understand the mechanisms that adaptive systems employ in achieving their ends.

Final Comment: Mind and Brain

My discussion, as well as Anderson's, has been carried on wholly at the level of symbolic information processes. Neither of us has mentioned, though we surely both believe, that thinking is implemented by a biological organ called the brain. Architecture is simply a high-level description of the brain in terms of its information-processing properties. To claim that architecture is more notation than substance is to make the same claim for the brain -- the fact that it supports adaptive behavior makes unnecessary any curiosity about how it operates.

Now such a claim is not wholly outrageous. In fact, a weaker form of the same claim is at the root of the idea that human thinking can be simulated (its processes as well as its outcomes) by computer. The exact ways in which neurons accomplish their functions is not important -- only their functional capabilities and the organization of these. Nothing else will show through to "behavior."

But what does show through is precisely what we have been calling "architecture." And for that reason architecture is by no means all notation; it has real substance in its effects on behavior. In my view, Anderson assigns too little weight to architecture (and by implication to strategies) as determinants of adaptive behavior.

References

Anderson, J. R., The place of cognitive architectures in a rational analysis, (this volume)

Anderson, J. R., *The adaptive character of thought* (unpublished manuscript)

Brunswik, Egon, *Perception and the representative design of psychological experiments*, Berkeley, CA: University of California Press, 1956.

Campbell, D. T., Evolutionary epistemology, in P. A. Schilpp (ed.), *The Philosophy of Karl Popper*, La Salle IL: Open Court, 1974.

Feigenbaum, E. A., and H. A. Simon, EPAM-like models of recognition and learning, *Cognitive Science* 8:305-336 (1984).

Gibson, J. J., *The ecological approach to visual perception*, Boston MA: Houghton-Mifflin, 1979.

Kotovsky, K., J. R. Hayes, and H. A. Simon, Why are some problems hard? *Cognitive Psychology* 17:248-292 (1985)

Kotovsky, K., and H. A. Simon, Why some problems are *really* hard, (unpublished manuscript)

Lumsden, C. J., and E. O. Wilson, *Genes, mind and culture*, Cambridge, MA: Harvard University Press, 1981.

Marr, David, *Vision*, San Francisco, CA: W. H. Freeman, 1982.

Newell, A., and H. A. Simon, *Human Problem Solving*, Englewood Cliffs, NJ: Prentice-Hall, 1972.

Simon, H. A., *Administrative Behavior*, New York, NY: Macmillan, 1947.

Simon H. A., *The Sciences of the Artificial*, 2nd edition, Cambridge, MA: MIT Press, 1978.

Simon, H. A. *Models of Thought*, New Haven, CT: Yale University Press, 1979.

Simon, H. A., *Models of Bounded Rationality*, Vol. 2, Cambridge, MA: MIT Press, 1982.

Simon, H. A., *Reason in Human Affairs*, Stanford, CA: Stanford University Press, 1986.